A quick review:

After reading 1 describing the Similarity ensemble approach (SEA), which is used to predict binding based on chemical similarity of ligands, it appeared that in the process of creating the statistical model, the reasoning behind some of the steps was lacking or not provided.

In creating the statistical model, the following outline was followed:

1. Sets of ligands “across logarithmic set size intervals in the range of 10 to 1,000 molecules” were populated randomly.
2. For all pairs of ligands, one from set A the other from set B, the Tanimoto coefficient score (Tc; ranging from 1 - maximal similarity, to 0 - minimal similarity) was calculated. The resulting sum of Tc scores (; raw score) was plotted vs. the number of calculations made (*CS*). CS equals the multiplication of the sizes of set A and B, which I will refer to as the comparison size.
3. Grouping by *CS* the mean and standard deviation (SD) of RS were calculated and a linear model was fit to *CS* against the mean while a non-linear model was fit for *CS* against SD.
4. For each comparison made normalized z-scores () were calculated as follows: .
5. A Tc cutoff of 0.57 was selected to force the resulting distribution to match an extreme value distribution (EVD).

Three issues emerged:

First, sampling was done “across logarithmic set size intervals in the range of 10 to 1,000 molecules”, which itself is unclear and although the minimum of CS is 100 (between two groups of size 10) the model is later being used to conclude the significance of results even when the *CS* is less than 100.

Second, SD was fitted using a non-linear regression and it is clear from the graph showing SD vs. *CS* that real SD values are getting further away from the predicted SD values as *CS* grows.

Third, it is not clear why the *ZS* distribution should be coerced into an EVD. The reasoning given in the original article [give reference] cites BLAST as showing that the EVD indicates biological significance, however in BLAST the underlying process is built in a way that maximizes the score (by finding the best alignment between two proteins).

Thus, we decided to derive a statistically robust model.

Methods:

All raw data used was provided by Keiser from a paper published in 2012 2. This includes: A mapping between protein targets and their ligands, ligand fingerprints, which were used to derive the statistical model, and drug fingerprints, which are not included in the ligand set, which were shown to bind to specific targets 2.

To try and resolve the emerging issues we repeated the original article’s [reference] sampling process in order to create a similar set of data to be used in the analysis step.

Sampling process:

1. Ligand sets of sizes ranging from 1 to 1,000 were randomly selected from a pool of 163,547 ligands.
2. For all pairs of ligands, L1 and L2, where one is from set A and the other from set B, the Tanimoto coefficient score [repetition] was calculated. The sum of Tc was saved, to the tuple (, , ).

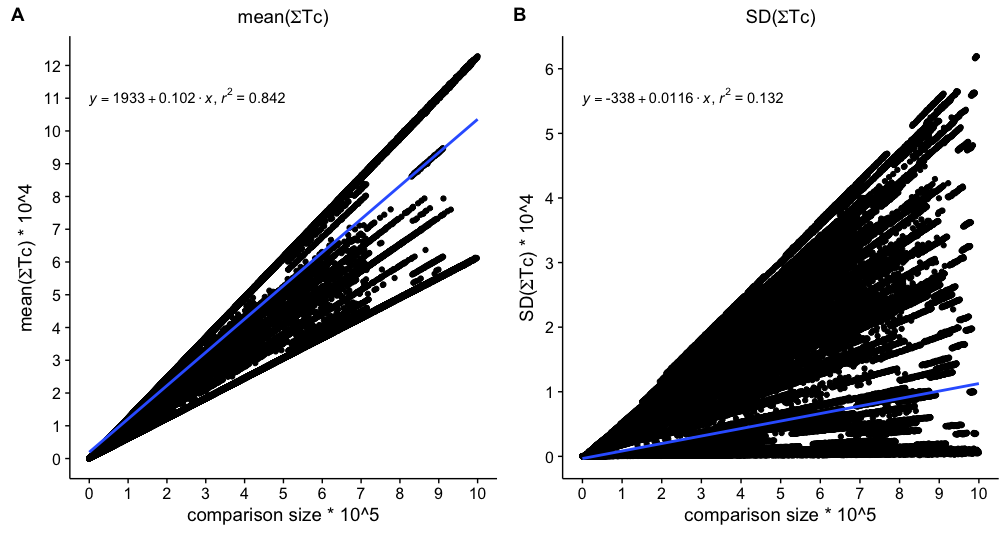
The sampling process was repeated 10 times.

Analysis done:

In SEA 1 sampling was done using “across logarithmic set size intervals in the range of 10 to 1,000 molecules” which resulted in set comparison sizes () ranging from 100 to. In that paper there were no single ligand sets in the sampling process. However, the same statistical model was later used to predict the statistical significance of the score when a single ligand is checked vs. another set of ligands, which sometimes resulted in a comparison size less than 100 (Keiser et al. 2012) [add to bibliography]. Thus, we examined if the same results would be achieved when using set sizes ranging from 1 to 1,000 (using set size interval of 1).

In order to replicated the sampling data produced by 1 when creating the SEA statistical model we calculated a mapping between (= CS) and .

The results are shown in Figure 1:



**1** Replicated sampling results. (**A**) Linear regression drawn in blue was fitted to the mean of the Tc scores sum plotted against CS (the number of calculations made). (**B**) Linear regression drawn in blue was fitted to the SD of the Tc scores sum plotted against CS.

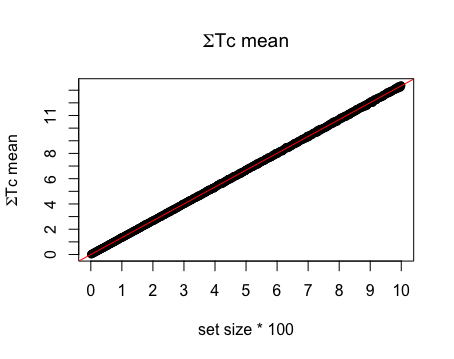
Fig. 1A shows the mean of for each comparison size. When fitting a linear model to mean, the model has an R-squared value of 0.8421 but a calculated mean square error (MSE) of 1.137e+08.

Fig. 1B shows the standard deviation (SD) of for each comparison size.

When fitting a linear model to SD, the model has an R-squared value of 0.1325 and a calculated MSE of 5.179e+07. It is easy to see that trying to fit any model to that data would not produce usable results. (In the paper describing SEA [Keiser 2007?] a nonlinear fit was computed). [The plot shows multiple linear lines. **Could you suggest why?** I’m not sure these are in fact multiple linear lines, I can try and create a model as I’ve done for the comparison of a single ligand and plot those lines on the two graphs.] [I do see lines and do not understand your reply. It is fine if you simply point out the fact that there are lines, and say that we do not understand why. Maybe Saharon will.]

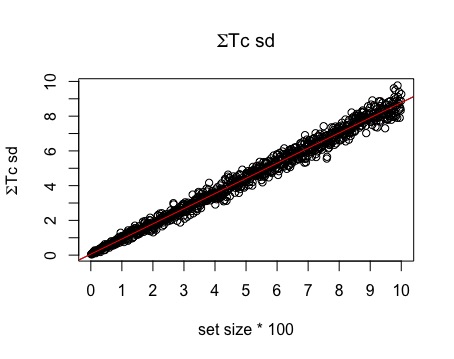
Next, to tackle issues two and three, we examined the possibility that the data above corresponds to a mixture. Specifically, that the results of comparisons of the same size may differ based on the sizes of the sets used. For example, a comparison of size 400 can be obtained by comparing one ligand to a set of 400 ligands, as well as by comparing two sets of 20 ligands each. We propose that the results of the model are based on the size of the smaller of the two sets. To test this, the data was grouped based on the size of the smaller set of the two used for the comparison.

The results shown below are for a small set of one ligand only. As, in practice, this is the most useful model.



**2** Linear regression drawn in red was fitted to the mean of the Tc scores sum plotted against CS.

We fitted a linear regression which gave an R-squared value of 0.999 and MSE of 0.07702.



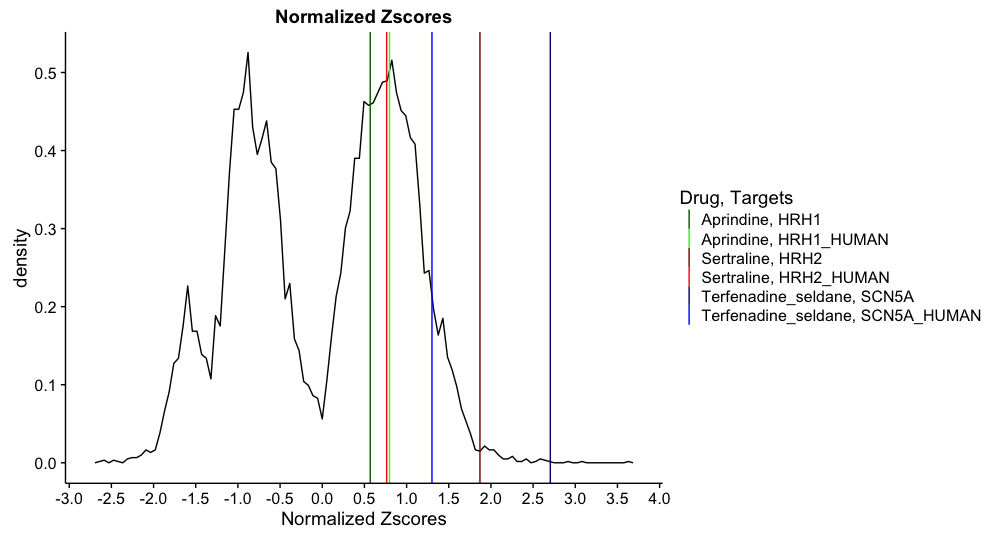
3. Linear regression drawn in red was fitted to the SD of the Tc scores sum plotted against CS.

We fitted a linear regression which gave an R-squared value of 0.9868 and MSE of 0.08477.

Evidently by taking into account the size of the smaller set we receive a better fit to the mean, and much better to the SD. Thus, it is interesting to see how the distribution of *ZS* changes using the revised statistical model. The ZS histogram in Fig. 4 suggests that we are dealing with a bimodal normal distribution, depicting two populations: The right most would correspond to ligands that bind to their targets, and the left most of ligands that do not.

In order to test this, we used Sertraline and other drugs and their targets that has been shown in-vitro by 2 to be positive predictions by the (original) SEA model. Each drug’s *ZS* was calculated against two target proteins groups, the first is human only, and the second is all organisms for which binding information is available. This is done since in 2 it is unclear whether the predictions were made against human proteins only or against all organisms for which binding information is available. Mapping the results into the distribution of Fig. 4 shows that all these binders appear on the right, in support of our suggestion.

We plan to ask Shoichet and Keiser to provide details of false positives and ambiguous results (predictions which do bind but at a high concentrations) and see where they appear in the histogram of Fig. 4.



**4**. Normalized z-scores density plot. Plotted as vertical lines are positive predictions made by SEA.

1. Keiser, M. J. *et al.* Relating protein pharmacology by ligand chemistry. *Nat Biotech* **25,** 197–206 (2007).

2. Lounkine, E. *et al.* Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486,** 361 (2012).